ViTopic – Topic Modeling on Images by Clustering of Vision Transformer Embeddings

—

By: Fausto J. German Jimenez

Project Homepage: https://github.com/faustotnc/vitopic/

Problem Statement & Motivation

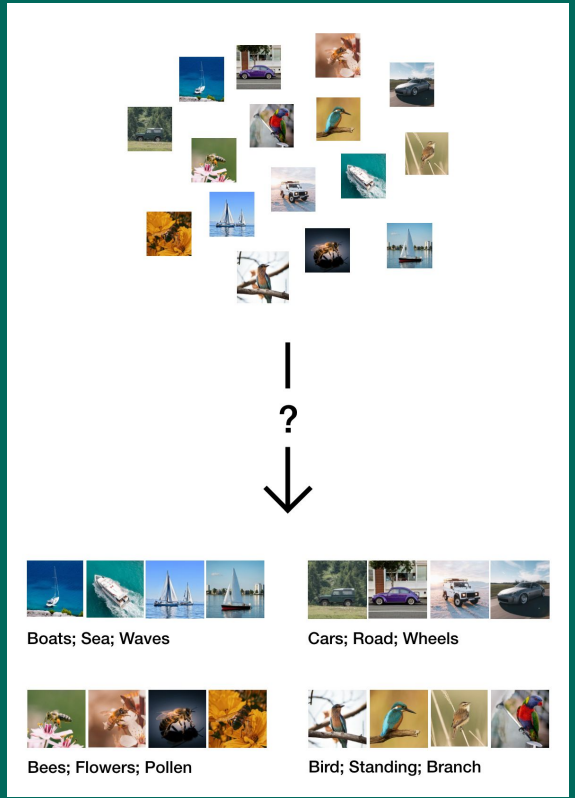
- Image data plays an important role in fields like social media analysis, medical research, and agriculture.
- Collecting, sorting, and clustering image data are resource-intensive tasks.
- Human intervention is still need to interpret the contents of groups of images.

Problem Statement

- How do we cluster a collection of unlabeled images and generate topic descriptors based on their visual context?

Why is this important?

- Automating every step of the process could drastically decrease labor costs and make analysing image data more accessible.



* Images from Unsplash

Related Work

Title: Contextual Visual Similarity

Authors: Xiaofang Wang, Kris M. Kitani and Martial Hebert

Year: 2016

Publisher: ArXiv

URL: https://arxiv.org/abs/1612.02534

Related Work

In their research the authors use triples of images for unsupervised attribute discovery, where they learn feature weights for each triplet such that the distance between two triplets is small if they are similar.

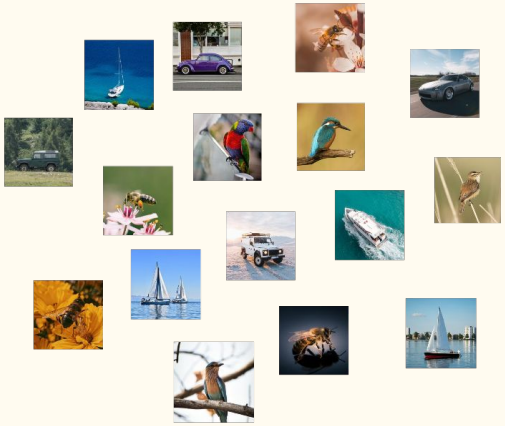
The main difference between their method and the methods I plan to use is that they require three images (A query image, a positive example, and a negative example) to define a contextualized similarity search criteria, while ViTopic computes similarity based on the Inverse Euclidean Distance between Vision Transformer Embeddings.

Other Related Work

This project combines the work from recent research on computer vision models and topics modeling techniques:

1. **Swin Transformer: Hierarchical vision transformer using shifted windows.**
   a. By Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Published to ArXiv on August 17, 2021.
   b. **Why this paper?** It describes Vision Transformers, which I plan to use to compute image embeddings for contextual similarity.
2. **Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation**
   a. By Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Published to ArXiv on February 15, 2022.
   b. **Why this paper?** It describes the model I plan to use for automated image-captioning.
3. **BERTopic: Neural topic modeling with a class-based TF-IDF procedure.**
   a. By Maarten Grootendorst. Published to ArXiv on March 11, 2022.
   b. **Why this paper?** It describes the pipeline I plan to use to create contextual topics.

Main Idea – Visually Explained



???

Boats; Sea; Waves

Bees; Flowers; Pollen

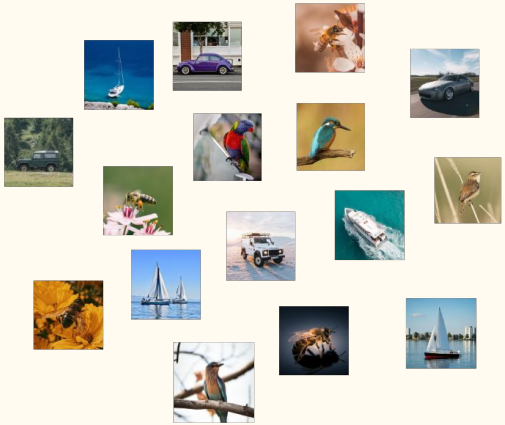
Cars; Road; Wheels

Bird; Standing; Branch

How do we cluster a collection of unlabeled images and generate topic descriptors based on their visual context?

* Images from Unsplash

Main Idea – Visually Explained



Boats; Sea; Waves

Bees; Flowers; Pollen

Cars; Road; Wheels

Bird; Standing; Branch

Each image is assigned a cluster (as determined by their transformer embeddings). Then, each image is described using an image captioning model, and topic descriptors are generated based on the frequency of words in the captions.

* Images from Unsplash

The Data

I will use the ImageNette dataset, which is a small subset of the ImageNet project.

We can easily access this subset using the Datasets library from the HuggingFace repository: https://huggingface.co/datasets/frgfm/imagenette

This dataset will make it possible to evaluate the performance of the system since it contains labeled images within a small number of classes.

https://paperswithcode.com/dataset/imagenet

The Pipeline (Methods)




microsoft/swin-base-patch4-window12-384-in22k

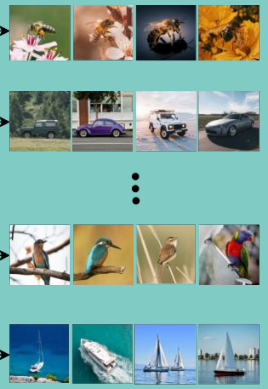
Vision Transformer Model

Vision Embeddings

$$d(\vec{v_1}, \vec{v_2}) = \frac{||\vec{v_1}||}{||\vec{v_1}|| + 2||\vec{v_1} - \vec{v_2}||}$$

Visual Similarity Search

Each row of this matrix is the similarity between one embedding and every other embedding in the dataset



Context-Guided Topic Model

Text Captions

BLIP Captioning Model

Salesforce/blip-image-captioning-large

UMAP Dimensionality Reduction

HDBSCAN Density-Based Clustering

Topic Cluster #1

Topic Cluster #2

Topic Cluster #n-1

Topic Cluster #n

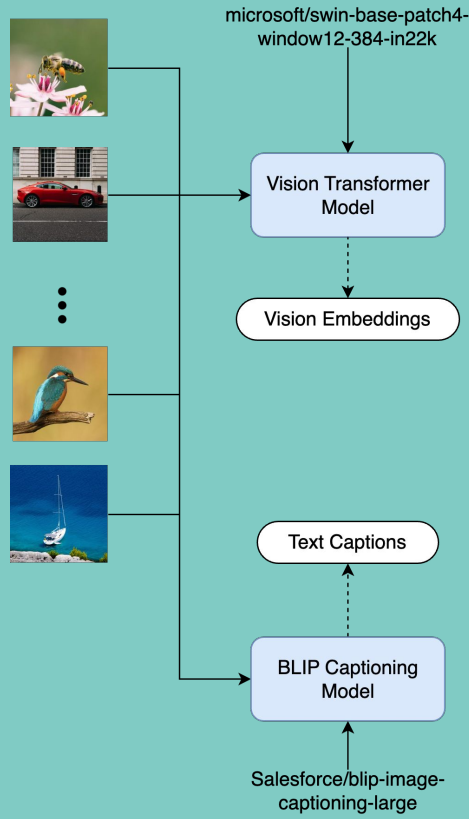
c-TF-IDF

Relevant Words In Cluster

Cluster-Based TF-IDF. This is different from regular TF-IDF in that the matrix is computed cluster-wise and not document-wise.

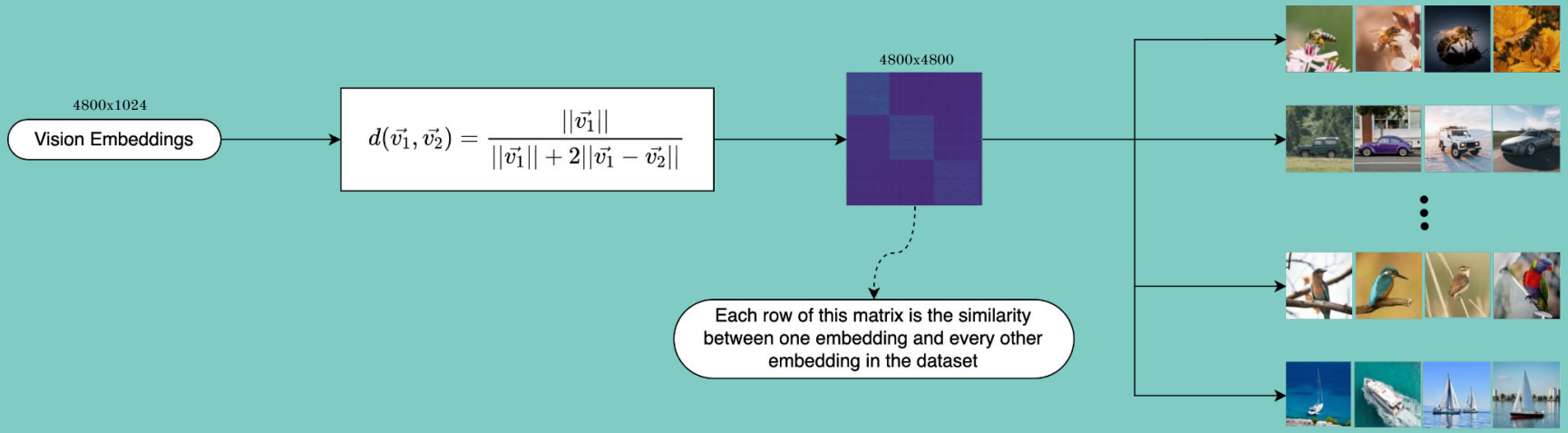
* Images from Unsplash

The Pipeline (Methods) – Embeddings & Captions



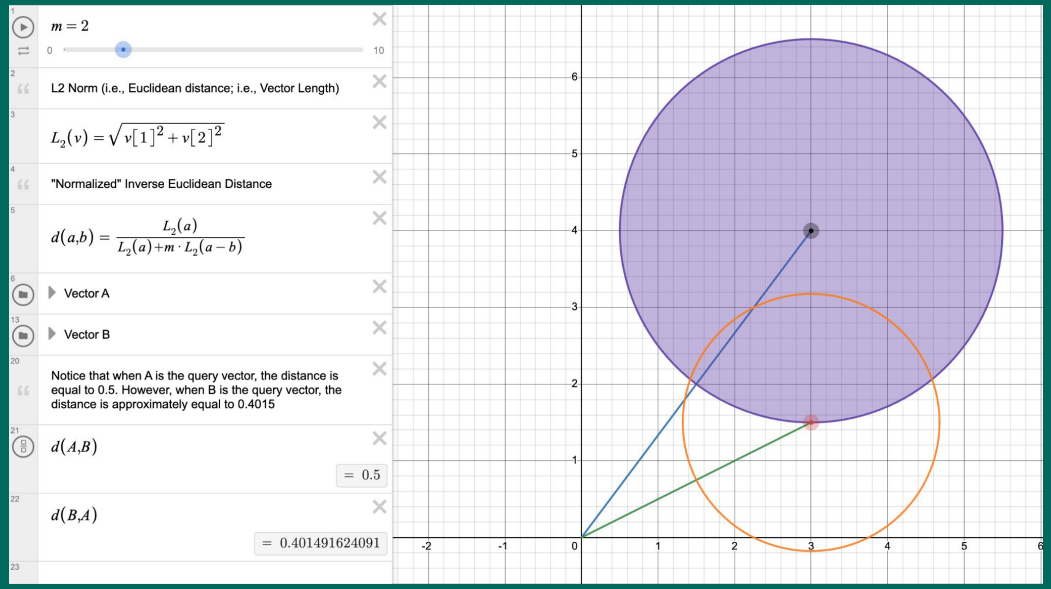
- Take the images and generate vision embeddings and captions
- SWIN Transformer Model for Vision Embeddings
- BLIP Model for Image Captioning

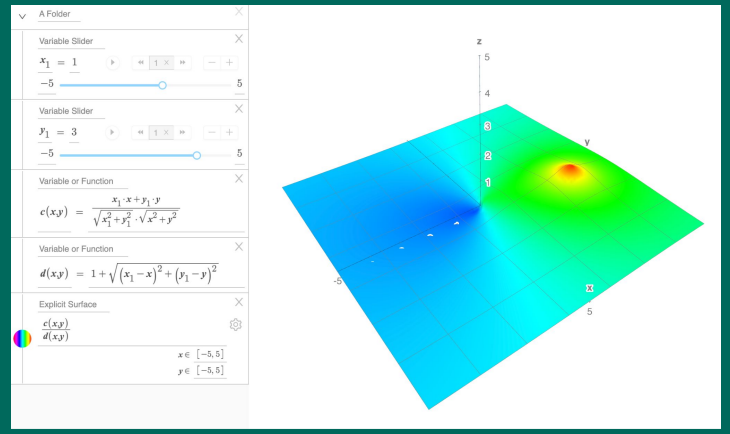
The Pipeline (Methods) – Visual Similarity Search



- Generate similarity matrix from vision embeddings
- Use Inverse Euclidean distance to compute similarity scores
- Sort entries by value

Side Note – Similarity Function & Variations





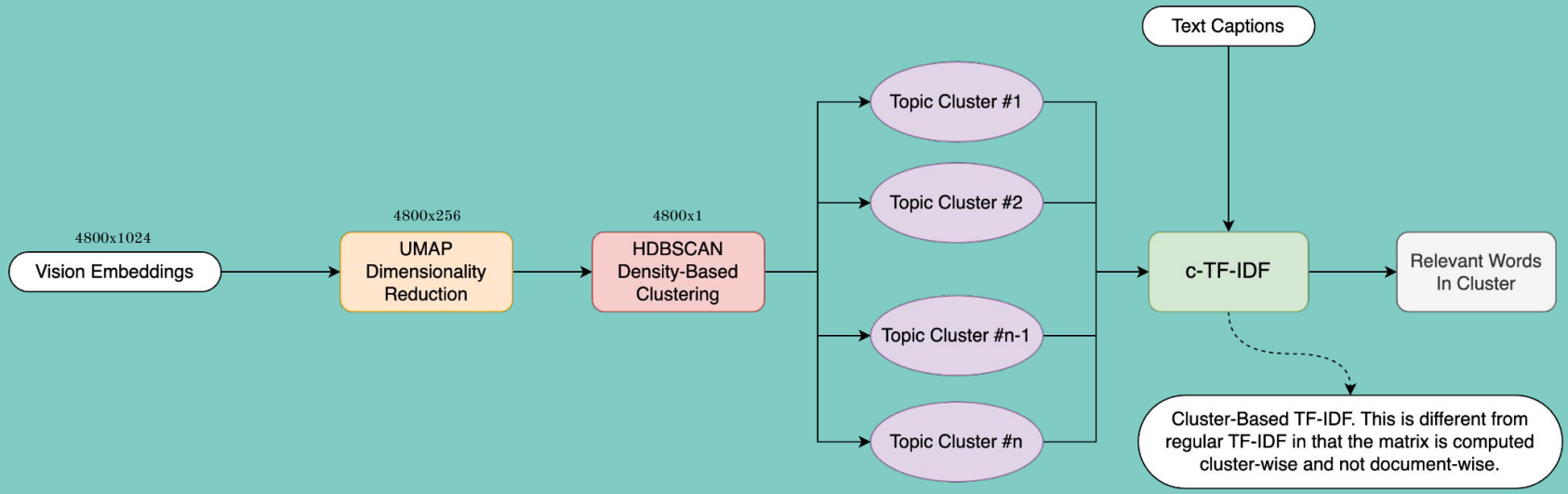
- "Normalized" inverse Euclidean distance
- Takes into account the length of each embedding vector

Why not Cosine Similarity?

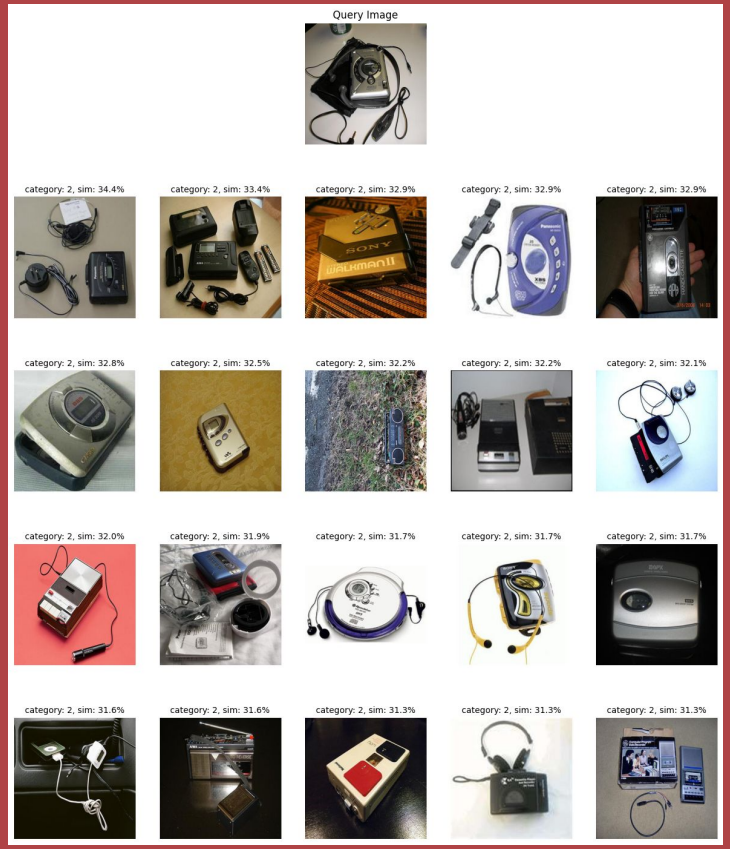
- Incorporating cosine similarity into the equation produces similar results
- Unlike text embeddings, image embeddings do not capture polarity very well

The Pipeline (Methods) – Context-Guided Topic Model

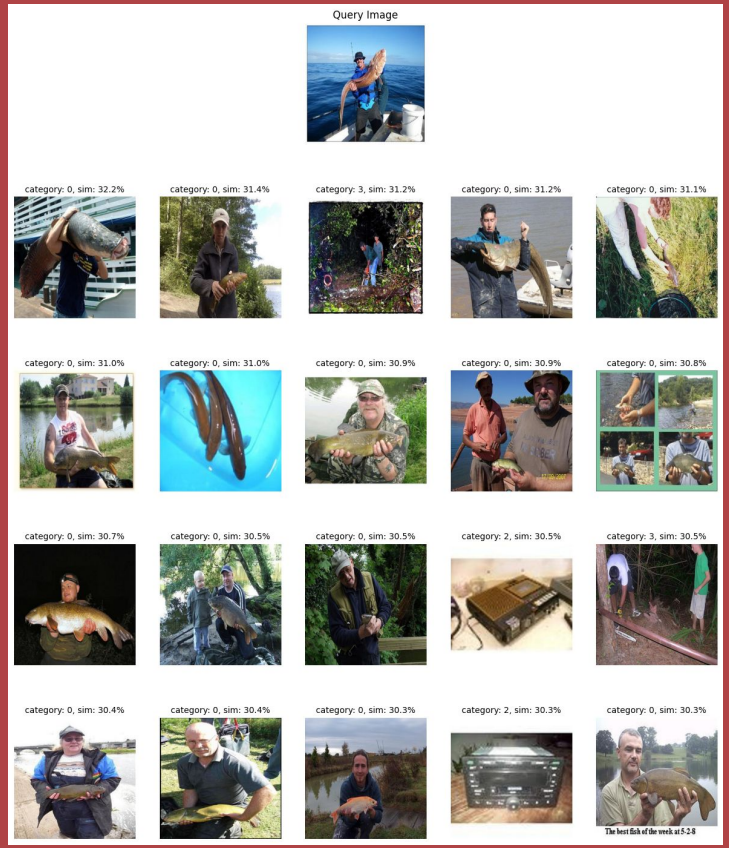
- Use vision embeddings and text captions to find topic descriptors (frequent words) guided by the contextual information of the images
- Use UMAP & HDBSCAN for clustering of vision embeddings
- Use the text captions and cluster information for topic discovery



Results – Visual Similarity Search

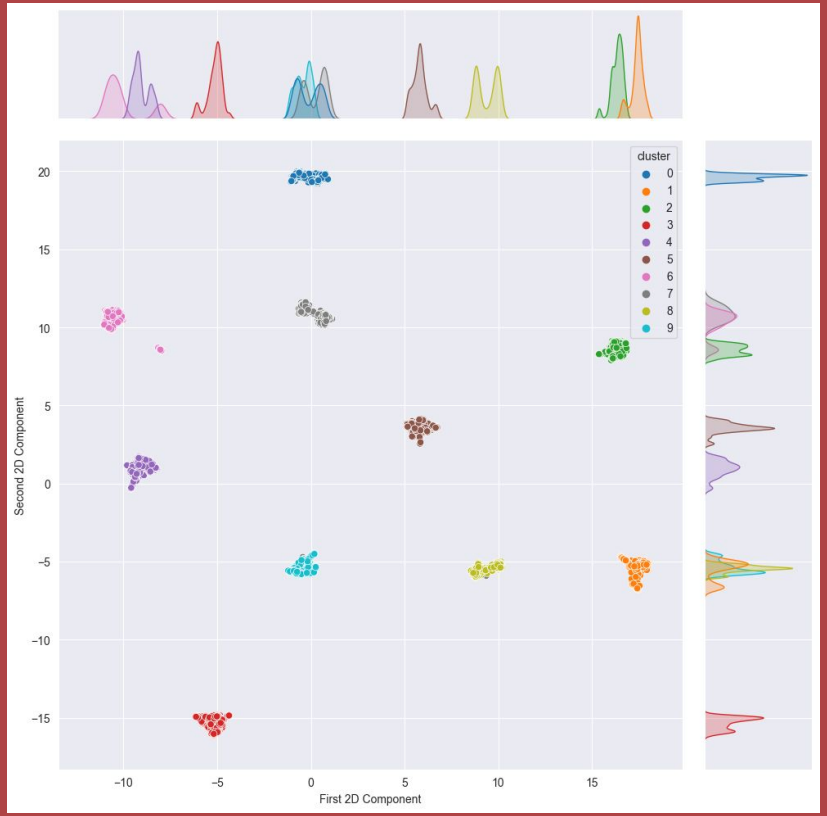


Results – Visual Similarity Search



It's not a perfect solution.

Results – Context-Guided Topic Model



	Topic #0		Topic #1		Topic #2		Topic #3	
1	golf	0.41285	fish	0.42141	dog	0.44577	truck	0.43295
2	ball	0.32763	holding	0.22051	laying	0.16123	araf	0.27578
3	close	0.18265	man	0.19609	grass	0.144	parked	0.21547
4	balls	0.17148	caught	0.17625	standing	0.1381	garbage	0.19853
5	grass	0.10705	hands	0.14238	sitting	0.12956	road	0.17481
6	green	0.10309	kneeling	0.13034	dogs	0.10371	street	0.17224
7	tee	0.08229	araffe	0.10046	floor	0.09872	driving	0.13885
8	field	0.06633	arafed	0.09745	mouth	0.09101	lot	0.09386
9	club	0.04815	net	0.09248	frisbee	0.09098	parking	0.08043
10	hole	0.04744	river	0.07655	puppy	0.08489	standing	0.0659

	Topic #4		Topic #5		Topic #6		Topic #7	
1	radio	0.33948	parachute	0.30839	church	0.3368	horn	0.23816
2	close	0.20243	flying	0.30473	steeple	0.20044	French	0.21051
3	player	0.17036	sky	0.24832	clock	0.1917	brass	0.20206
4	table	0.12594	araffe	0.17634	tower	0.18597	playing	0.18343
5	stereo	0.11957	air	0.15293	view	0.17128	instrument	0.15722
6	cassette	0.11668	person	0.1179	arafed	0.16009	instruments	0.1059
7	remote	0.09721	kite	0.1165	building	0.10356	sitting	0.08779
8	cd	0.07453	people	0.0876	large	0.09353	silver	0.08279
9	sitting	0.07121	flag	0.0816	cross	0.07359	room	0.08119
10	speakers	0.06938	parasailing	0.0775	white	0.07264	woman	0.07964

	Topic #8		Topic #9	
1	chainsaw	0.37854	gas	0.43259
2	tree	0.17521	pump	0.33474
3	cutting	0.13872	station	0.12938
4	man	0.12494	pumps	0.10886
5	handle	0.09879	sitting	0.10357
6	arafed	0.09543	sign	0.1013
7	cut	0.091	old	0.09499
8	using	0.08525	red	0.08703
9	dew	0.07936	building	0.076
10	close	0.06073	arafed	0.07219

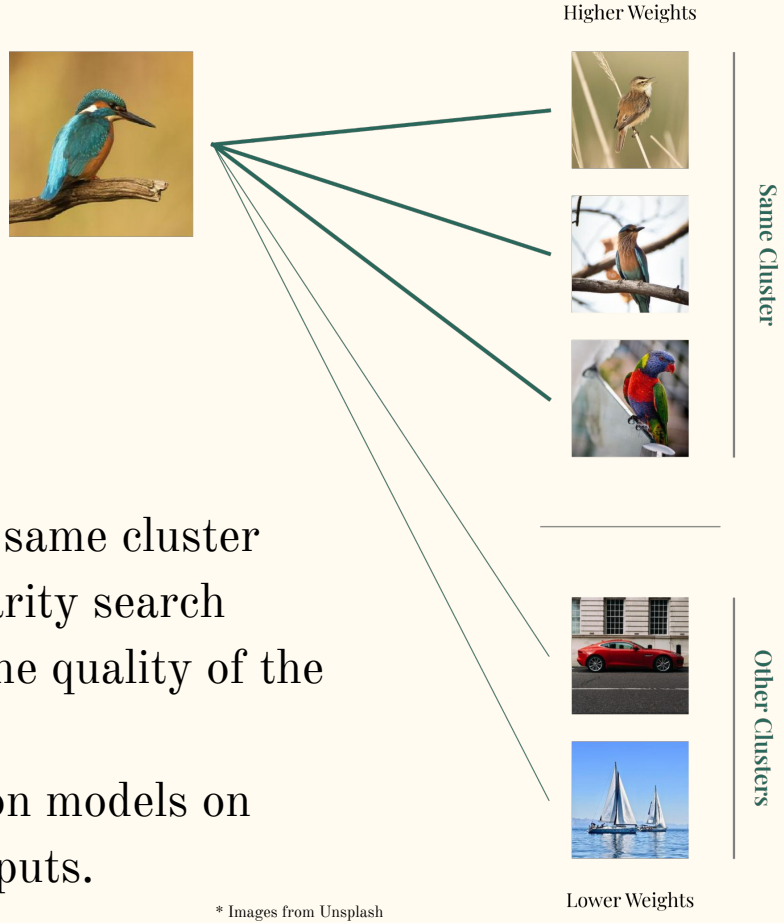
- There is some overlap between topics #1 and #8
- This relates to the results we saw in the visual similarity search

Conclusion & Future Work

- Using vision transformer embeddings with image captioning is a viable method for clustering images based on context topics

Potential Improvements

- Prioritizing neighboring vectors within the same cluster could improve the result of the visual similarity search
- Using newer vision models could improve the quality of the embeddings
- Using transfer learning to re-train the vision models on specific tasks could greatly improve the outputs.



* Images from Unsplash

References

- Grootendorst, M. (2022, March 11). BERTopic: *Neural topic modeling with a class-based TF-IDF procedure*. arXiv.org. Retrieved March 17, 2023, from https://arxiv.org/abs/2203.05794
- Li, J., Li, D., Xiong, C., & Hoi, S. (2022, February 15). *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*. arXiv.org. Retrieved March 17, 2023, from https://arxiv.org/abs/2201.12086
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021, August 17). *Swin Transformer: Hierarchical vision transformer using shifted windows*. arXiv.org. Retrieved March 17, 2023, from https://arxiv.org/abs/2103.14030
- Wang, X., Kitani, K. M., & Hebert, M. (2016, December 8). *Contextual visual similarity. arXiv.org*. Retrieved April 1, 2023, from https://arxiv.org/abs/1612.02534